

A Generative Modeling Approach to Calibrated Predictions: A Use Case on Menstrual Cycle Length Prediction

Iñigo Urteaga

INIGO.URTEAGA@COLUMBIA.EDU

*Department of Applied Physics and Applied Mathematics, Data Science Institute
Columbia University, New York, NY, USA*

Kathy Li

KATHY.LI@COLUMBIA.EDU

*Department of Applied Physics and Applied Mathematics, Data Science Institute
Columbia University, New York, NY, USA*

Amanda Shea

AMANDA.SHEA@BIOWINK.COM

Clue by BioWink, Adalbertstraße 7-8, 10999 Berlin, Germany

Virginia J. Vitzthum

VITZTHUM@INDIANA.EDU

*Kinsey Institute & Department of Anthropology
Indiana University, Bloomington, IN, USA*

Chris H. Wiggins

CHRIS.WIGGINS@COLUMBIA.EDU

*Department of Applied Physics and Applied Mathematics, Data Science Institute
Columbia University, New York, NY, USA*

Noémie Elhadad

NOEMIE.ELHADAD@COLUMBIA.EDU

*Department of Biomedical Informatics, Data Science Institute
Columbia University, New York, NY, USA*

Editor: Editor's name

Abstract

We explore how to quantify uncertainty when designing predictive models for healthcare to provide well-calibrated results. Uncertainty quantification and calibration are critical in medicine, as one must not only accommodate the variability of the underlying physiology, but adjust to the uncertain data collection and reporting process. This occurs not only on the context of electronic health records (i.e., the clinical documentation process), but on mobile health as well (i.e., user specific self-tracking patterns must be accounted for). In this work, we show that accurate uncertainty estimation is directly relevant to an important health application: the prediction of menstrual cycle length, based on self-tracked information. We take advantage of a flexible generative model that accommodates under-dispersed distributions via two degrees of freedom to fit the mean and variance of the observed cycle lengths. From a machine learning perspective, our work showcases how flexible generative models can not only provide state-of-the-art predictive accuracy, but enable well-calibrated predictions. From a healthcare perspective, we demonstrate that with flexible generative models, not only can we accommodate the idiosyncrasies of mobile health data, but we can also adjust the predictive uncertainty to per-user cycle length patterns. We evaluate the proposed model in real-world cycle length data collected by one of the most popular menstrual trackers worldwide, and demonstrate how the proposed generative model provides accurate and well-calibrated cycle length predictions. Providing meaningful, less uncertain cycle length predictions is beneficial for menstrual health researchers, mobile health users and developers, as it may help design more usable mobile health solutions.

1. Introduction

One of the primary challenges in predictive modeling for healthcare pertains to handling the uncertainty of both the task and the data at hand, as well as ensuring calibration of model output (Han et al., 2011; Rogers and Walker, 2016; Chen et al., 2020). Because users of these models make decisions—that have health and ethical implications (Gillon, 1994; Siebert, 2003)—based on such predictive outputs, it is critical to ensure users can assess the confidence of a model in its predictions.

In machine learning predictions, different types of uncertainty are entangled. The uncertainty, given a finite amount of data, of a machine learning technique can be captured with the **statistical characterization of its predictions**, i.e., via the predictive distribution $p(y|x)$ of the model output y given the input features x . Entangled in this predictive distribution are both *aleatoric* and *epistemic* uncertainties. The former denotes the randomness inherent in the data generating process, i.e., the observed data (e.g., collected features and observed outcomes). The latter—also known as model uncertainty—reflects the uncertainty of a model’s appropriateness to fit the underlying data generating mechanism.

A goal of statistical machine learning is to devise suitable measures of the uncertainty associated with model predictions (Gneiting et al., 2007). Predictions are probabilistic in nature, taking the form of probability distributions over future events of interest (Dawid, 1984). In a Bayesian view of predictive modeling, the outcome Y , the input features X , and the parameters of a model are viewed as random variables¹. The distributional assumptions over the model class and the uncertainty over parameters can be characterized with priors and incorporated into **the predictive distribution** via marginalization of such (parametric) model uncertainties, i.e., $p(Y|x) = \int_{\theta} p(Y|x, \theta)p(\theta|\mathcal{D})d\theta$, where \mathcal{D} refers to previously observed data.

Within this probabilistic view of prediction, where the predictive distribution characterizes all the uncertainty in the outcome of interest Y , a model’s calibration is a crucial aspect. **Calibration** refers to the statistical consistency between the distributional forecasts (i.e., the predictive distribution) and the true observations (i.e., the data). As originally argued by Dawid (1984) and many after (Diebold and Mariano, 2002), the predictive posterior must be assessed on the basis of the predicted-observation pairs.

A statistical method is calibrated when, for all the observed examples x for which it predicts an outcome $Y = y$ with probability $p(Y = y|x) = p'$, the proportion (frequency) of real examples observed for outcome y is equal to p' , across all values of $P(Y = y|x)$. In this sense, calibration is inherently frequentist, but Bayesian views of calibration have also been argued for (Dawid, 1982). Essentially, calibration is concerned with measuring the over-confidence and under-confidence of a statistical model. As such, it helps assess the extent to which a user can trust the model’s predicted outcome probabilities.

In healthcare, the role of uncertainty estimation and model calibration is gaining momentum (Vach, 2013; Calster et al., 2016; Alba et al., 2017; Stevens and Poppe, 2020; Goldstein et al., 2021), partly due to the increase in popularity of deep learning based approaches (Rajkomar et al., 2018). Recent evidence suggests that deep-learning approaches lack calibrated outputs (Guo et al., 2017; Nixon et al., 2019), even if several techniques have been proposed to assess and fix this gap (Dusenberry et al., 2020; Kwon et al., 2020).

1. We capitalize random variables, we denote their realizations in lower-case.

Most of the healthcare model calibration work so far is on classification tasks in the context of clinical, electronic health record data. On the contrary, the use case for this work is **menstrual cycle length prediction from mobile health data**. Based on a user’s self-tracked data from a period tracking app, we aim at forecasting their upcoming cycle length, i.e., the date of their next period. This use case differs from previous work in quantifying uncertainty: (i) we target regression (i.e., next cycle length prediction) rather than classification, and (ii) we leverage mobile health (mHealth) data —subject to self-tracking artifacts, like varying adherence to tracking.

When characterizing and predicting menstrual patterns based on mHealth data, the relevance of uncertainty quantification is two-fold. On the one hand, self-tracked data from mHealth apps reflects both physiological menstrual patterns and user engagement dynamics (Li et al., 2020). Therefore, one must disentangle the uncertainty of the physiological process (i.e., the menstrual cycle) from the uncertainty on the observed data (i.e., whether users track their period). On the other, uncertain predictions in mHealth often result in non-actionable insights, e.g., “*your next period will occur within the next two weeks*” (Orchard, 2019; Fox and Epstein, 2020).

We hereby operate within a generative modeling framework, and leverage advances in the statistical characterization of complex distributions to accommodate both self-tracking artifacts and cycle length variability (Li et al., 2020). Precisely, we take advantage of probabilistic machine learning (Chen et al., 2020) and devise a flexible generative model that can accommodate the uncertainties of the task at hand and address predictive calibration directly. We demonstrate how to overcome the over-dispersion of Poisson distributed predictions by proposing a Generalized Poisson based model that provides accurate and better calibrated individualized cycle length predictions. Less uncertain cycle length predictions are intrinsically beneficial, and we hypothesize they may also help design better menstrual mHealth solutions, ultimately increasing their usability.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work contributes to machine learning in the context of healthcare by first proposing a flexible generative model to provide accurate and well-calibrated predictions. We demonstrate that our model outperforms black-box neural network and state-of-the-art alternatives, by providing interpretable, accurate and well-calibrated predictions. Armed with well-calibrated predictions, users can trust and act upon predictions with more confidence. Second, we argue for the use of probabilistic modeling in healthcare (Chen et al., 2020), since interpretability, accuracy and uncertainty quantification are critical in the medical domain. More broadly, we advocate for the machine learning in healthcare community to not only focus on point estimate based metrics, but to incorporate the calibration tool-set presented here into the evaluation pipeline.

2. Related Work

Calibration in Predictive Models. A key endeavor of predictive modeling is to provide forecasts that appropriately quantify uncertainty (Gneiting et al., 2007). In healthcare, because of the practical and ethical costs of incorrect and over-confident predictions, there

is great value in assessing not only the predictive ability of a given model, but in measuring its uncertainty as well (e.g., by comparing, evaluating, and ranking competing methods).

A probabilistic view of prediction tasks takes the form of predictive posterior densities, and the challenge in evaluating them lies in the dichotomy between comparing predictive probability distributions with observations that are real (or discrete) valued. In general, calibration (the statistical consistency between the distributional forecasts and the observations) and sharpness (the concentration of the predictive distributions) are two key metrics in evaluating predictive posteriors. Many tools for checking calibration and sharpness have been proposed, some based on visual representations and others based on scoring rules (Gneiting et al., 2007). Scoring rules provide summary measures for the evaluation of probabilistic forecasts that assign a numerical score based on the predictive distribution and on the events that materialize.

Calibration in classification tasks (e.g., assessing the risk of a discrete set of events, like in disease prediction) implies transforming classifier scores into class membership probabilities. For these categorical predictive tasks, the expected calibration error (ECE) has become popular (Naeini et al., 2015; Guo et al., 2017; Dusenberry et al., 2020). It is a tractable way to approximate the calibration of a model given a finite dataset, although subject to certain limitations (Nixon et al., 2019). Characterizing probabilistic predictions of continuous variables is fundamentally different from calibrating categorical and binary variable predictions. We refer to (Gneiting and Raftery, 2007), where a theoretically grounded review of scoring rules for density, quantile and interval forecasts is provided.

The increase in popularity of deep learning (Goodfellow et al., 2016) has resulted in a scrutiny of the uncertainty and calibration performance of these techniques (Malinin and Gales, 2018; Yao et al., 2019). To capture uncertainty, existing state-of-the-art neural network based approaches often make use of ensemble, batch-norm or dropout techniques, yet have been often found to be miscalibrated (Guo et al., 2017; Nixon et al., 2019). Although various post-processing calibration methods have been proposed (Niculescu-Mizil and Caruana, 2005), calibration within deep learning is still a concern.

On the one hand, some propose to decouple training for good predictive accuracy from its calibration (Song et al., 2019), while others address calibration within-training via alternate loss functions (Avati et al., 2020). These approaches are not guaranteed to appropriately balance prediction and calibration. On the other, an alternative is to consider Bayesian deep learning (Wilson, 2020). However, there still remain many questions regarding the accuracy of the computed Bayesian posteriors (Wenzel et al., 2020), specially so when approximate inference is used (Foong et al., 2019).

The investigation of model uncertainty and calibration within the medical domain is also gaining attention, partly due to the rise of deep learning in healthcare (Dusenberry et al., 2020; Stevens and Poppe, 2020; Goldstein et al., 2021). Dusenberry et al. (2020) examined neural network methods to capture model uncertainty in the context of electronic health records (EHR), and acknowledged that there is still plenty of work to do on devising methods that reduce model uncertainty at both training and prediction time. In the work we present here, not only the healthcare context is different (mHealth data, instead of EHR), but our predictive task distinct (regression versus classification).

Menstrual Prediction from mHealth Data. Period tracking apps are some of the most popular smartphone apps (Wartella et al., 2016), and prediction of next period date is one of the most required feature from these app users (Epstein et al., 2017). Recent research has shown that while machine learning methods show promise, predicting cycle length is a challenging task (Pierson et al., 2018; Li et al., 2021), sometimes at the expense of a successful user-app interaction (Fox and Epstein, 2020).

First, menstruation is a complex process with inherent variation and uncertainty within and across individuals (Treloar et al., 1967; Chiazze et al., 1968; Ferrell et al., 2005; Vitzthum, 2009; Harlow et al., 2012). The analysis of massive data from menstrual tracking apps confirmed that there are indeed wide variations in cycle length across menstruators, as well as within longitudinal cycle lengths of the same individual (Symul et al., 2018; Li et al., 2020; Soumpasis et al., 2020). Second, self-tracking data also comprises uncertainty, as the tracking behavior of app users is varied across individuals and in time (Urteaga et al., 2020; Li et al., 2021).

In this work, starting from a state-of-the-art model for cycle length prediction by Li et al. (2021), we leverage flexible generative modeling to capture the intrinsic uncertainty of both the underlying physiological menstrual process and self-tracking data, and explore the value of calibrated predictions.

3. Methods

We hereby propose a hierarchical, Generalized Poisson-based generative model² for cycle lengths self-tracked via mHealth that (i) accounts for when individuals may forget to self-track their period, (ii) pulls population-level information and learns individualized cycle length patterns and self-tracking propensities, and (iii) enables per-individual cycle length uncertainty quantification that results in well-calibrated predictive posteriors.

It is a generative, hierarchical model (see Figure 1 and the generative process description in Section 3.1), in that all per-individual parameters are drawn from the same population level distribution, which allows for the incorporation of global menstrual pattern knowledge (via informative priors) and pulling cycle information across individuals (to learn population-level hyperparameters).

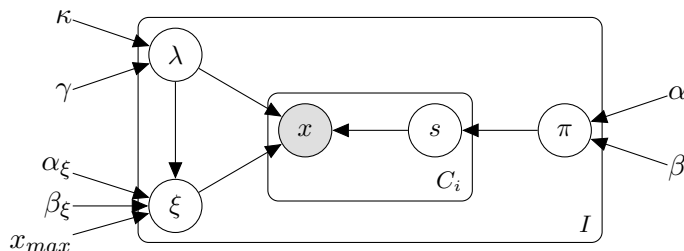


Figure 1: Probabilistic graphical model of the hierarchical Generalized Poisson model with a latent skipped cycle s variable to accommodate self-tracking artifacts.

2. A Python implementation of the proposed model is available in the public GitHub repository https://github.com/iurteaga/menstrual_cycle_analysis.

We take advantage of the Generalized Poisson distribution (Consul and Famoye, 2006) for its specific parameterization that enables us to capture the uncertainty of the observed cycle lengths (its posterior can be under- or over- dispersed) and provide calibrated predictions.

The Generalized Poisson is a distribution that belongs to the class of Lagrangian distributions over discrete, non-negative integers, with parameters λ and ξ that are independent (Consul, 1989; Consul and Famoye, 2006). The probability density function (pdf) of a Generalized Poisson, hereby denoted as $\mathcal{GP}(\cdot)$ follows

$$\mathcal{GP}(x|\lambda, \xi) = \begin{cases} \lambda(\lambda + \xi x)^{x-1} \frac{e^{-\lambda - \xi x}}{x!}, & \lambda > 0, x = 0, 1, \dots, \\ 0, & \text{for } x > x_{max} \text{ if } \xi < 0, \end{cases} \quad (1)$$

where limits on λ and ξ are imposed to ensure that there are at least five classes with nonzero probability (Consul, 1989). The first two moments³ of a $\mathcal{GP}(x|\lambda, \xi)$ follow $\mu_x = \frac{\lambda}{(1-\xi)}$ and $\sigma_x^2 = \frac{\lambda}{(1-\xi)^3}$, where we observe how the two independent parameters provide two degrees of freedom to fit the mean and the variance, separately. The Generalized Poisson can be over- or under-dispersed, depending on the value of ξ (when $\xi = 0$, we recover the Poisson distribution). For $\xi < 0$, the Generalized Poisson is under-dispersed (in comparison to a Poisson distribution) and it can be truncated to a maximum value x_{max} of x , requiring an additional normalizing factor $Z_{\mathcal{GP}(\lambda, \xi, x_{max})}$, see details in Section A.1 of the Appendix.

3.1. The hierarchical, generative process for observed cycle lengths

The proposed model, depicted in Figure 1, is a generative process with the following random (observed and unobserved) variables and parameters:

- The observed variables are the cycle lengths x_{i,c_i} , with $c_i = \{1, \dots, C_i\}$ cycle lengths for each individual $i = \{1, \dots, I\}$. We denote with s_{i,c_i} the (latent) number of skipped (not-reported) cycles, with $c_i = \{1, \dots, C_i\}$ cycle lengths for each individual $i = \{1, \dots, I\}$.
- λ_i and ξ_i denote the Generalized Poisson parameters for each individual $i = \{1, \dots, I\}$; π_i are the per-individual $i = \{1, \dots, I\}$ probability parameters of skipping a cycle.
- κ, γ are the population-level hyperparameters of a Gamma distribution prior over the λ_i ; α_ξ, β_ξ the hyperparameters of a Beta distribution prior over the ξ_i ; and α, β the hyperparameters of a Beta distribution prior over the skipping probabilities π_i .

We now summarize the generative process of the proposed probabilistic model. First, one draws individual cycle length and self-tracking probability parameters from the population level distributions: i.e., each individual's $\lambda_i \sim p(\lambda|\kappa, \gamma) = \mathcal{G}(\lambda|\kappa, \gamma)$ and $\xi_i \sim \max\{-1, \frac{-\lambda_i}{x_{max}}\} + (1 - \max\{-1, \frac{-\lambda_i}{x_{max}}\})\mathcal{B}(\xi|\alpha_\xi, \beta_\xi)$ parameters, and the the probability of each individual forgetting to track a period $\pi_i \sim p(\pi|\alpha, \beta) = \mathcal{B}(\pi|\alpha, \beta)$ (all distributional details are provided in Section A.2 of the Appendix). Given per-individual parameters λ_i, ξ_i, π_i , then the number of cycles a user forgets to track is drawn from a Truncated Geometric distribution with parameter π_i , i.e., $s_{i,c_i} \sim p(s|\pi_i) = \frac{\pi_i^s(1-\pi_i)}{\sum_{s=0}^{s_{max}} \pi_i^s(1-\pi_i)}$. Finally, the observed cycle length for each user i is drawn from a Generalized Poisson distribution, conditioned on the number of skipped cycles, i.e., $x_{i,c_i} \sim p(x|\lambda_i, \xi_i, s_{i,c_i}) = \mathcal{GP}(x|(s_{i,c_i} + 1)\lambda_i, \xi_i)$.

3. Other moments of interest can be computed in closed form, see (Consul and Famoye, 2006) for a full characterization of this distribution.

3.2. The proposed model’s distributions of interest

There are two distributions that are critical for the purpose of this study: (i) the joint, over all individuals’ marginalized data (log)-likelihood, and (ii) each individual’s cycle length predictive posterior. As explained in the introduction, a Bayesian view of predictive modeling requires the derivation of the predictive distribution via marginalization of all the inherent (parametric) model uncertainties:

$$p(Y|x) = \int_{\theta} p(Y|x, \theta) p(\theta|\mathcal{D}) d\theta. \quad (2)$$

Here, we aim at marginalizing all the parameters of the model, based on the prior distribution assumptions described in Section 3.1. In our proposed model, the latent parameters θ contain both the unobserved per-cycle skips s_{i,c_i} and the per-individual parameters λ_i , ξ_i and π_i , i.e., $\theta = \{s_{i,c_i}, \lambda_i, \xi_i, \pi_i\}$. Similarly, and to reduce clutter, we denote with Θ all the hyperparameters of our model: $\Theta = \{\kappa, \gamma, \alpha_{\xi}, \beta_{\xi}, \alpha, \beta\}$.

We clarify our notation here, where we denote a set of cycle length observations for a given individual i with $X_i = (x_{i,1}, \dots, x_{i,C_i})^{\top} \in \mathbb{R}^{C_i}$, and the set of cycle length observations for all individuals $i = \{1, \dots, I\}$ in the population are denoted with $X = (X_1^{\top}, \dots, X_I^{\top})^{\top} \in \mathbb{R}^{I \times C_I}$, where $C_I = \max C_i, \forall i$. Similarly, the set of latent skipped cycle variables for a given individual i is denoted with $S_i = (s_{i,1}, \dots, s_{i,C_i})^{\top} \in \mathbb{R}^{C_i}$, and the set of all latent skipped cycle variables for all individuals $i = \{1, \dots, I\}$ in the population are denoted with $S = (S_1^{\top}, \dots, S_I^{\top})^{\top} \in \mathbb{R}^{I \times C_I}$, where $C_I = \max C_i, \forall i$.

3.2.1. MARGINALIZED JOINT DATA LIKELIHOOD

The population level data likelihood, with marginalized parameters, can not be derived in closed form. Instead, we resort to a hybrid approach, where we analytically marginalize per-individual skipped cycles s_{i,c_i} and use Monte Carlo to marginalize the per-individual parameters λ_i , ξ_i and π_i . The resulting marginalized joint data likelihood follows

$$p(X|\Theta) = \prod_{i=1}^I \prod_{c_i=1}^{C_i} p(x_{i,c_i}|\Theta) \approx \prod_{i=1}^I \frac{1}{M} \sum_{m=0}^M p(X_i|\lambda_i^{(m)}, \xi_i^{(m)}, \pi_i^{(m)}), \quad (3)$$

where the per-user joint likelihood is marginalized over the skipped cycles s_{i,c_i} , i.e.,

$$p(X_i|\lambda_i^{(m)}, \xi_i^{(m)}, \pi_i^{(m)}) = \prod_{c_i=1}^{C_i} \sum_{s_{i,c_i}=0}^{s_{max}} p(x_{i,c_i}|s_{i,c_i}, \lambda_i^{(m)}, \xi_i^{(m)}) p(s_{i,c_i}|\pi_i^{(m)}), \quad (4)$$

and evaluated with Monte Carlo parameters $\lambda_i^{(m)} \sim p(\lambda|\kappa, \gamma)$, $\xi_i^{(m)} \sim p(\xi|\lambda_i^{(m)}, \alpha_{\xi}, \beta_{\xi})$, and $\pi_i^{(m)} \sim p(\pi|\alpha, \beta)$ drawn from their respective prior/posterior distributions (corresponding Equations (11), (12), and (13) are provided in Section A.2 of the Appendix).

The joint data likelihood is key for our training procedure, and determines the computational complexity of the fitting procedure. Given a dataset of C_i cycle lengths for $i = \{1, \dots, I\}$ users, we perform hyperparameter inference via type-II maximum likelihood estimation; that is, we find the hyperparameters $\hat{\Theta}$ that maximize the data log-likelihood as provided in Equation (3), i.e., $\hat{\Theta} = \text{argmax}_{\Theta} [\ln(p(X|\Theta))]$.

After the training procedure, the hyperparameters $\Theta = \{\kappa, \gamma, \alpha_\xi, \beta_\xi, \alpha, \beta\}$ used for drawing the Monte Carlo parameters in Equation (3) will be replaced with the learned population level hyperparameters $\hat{\Theta}$.

We note that the hierarchical nature of the proposed model enables distributed learning, with not only computational, but also privacy benefits: mHealth users do not need to share their data (they can locally compute their individualized predictions), and only need to share per-user data log-likelihood estimates for population-level hyperparameter inference (see Section A.6 of the Appendix for a more detailed discussion).

3.2.2. CYCLE LENGTH PREDICTIVE POSTERIOR

We derive the marginalized predictive posterior of the next cycle length $x_{i,c_{new}}$ after observing per-user cycle lengths X_i ,

$$p(x_{i,c_{new}}|X_i, \Theta) = \frac{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} p(x_{i,c_{new}}|\lambda_i, \xi_i, \pi_i)p(X_i|\lambda_i, \xi_i, \pi_i)p(\lambda_i, \xi_i, \pi_i|\Theta)d\lambda_i d\xi_i d\pi_i}{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} p(X_i|\lambda_i, \xi_i, \pi_i)p(\lambda_i, \xi_i, \pi_i|\Theta)d\lambda_i d\xi_i d\pi_i}, \quad (5)$$

for which we need to compute $p(X_i|\lambda_i, \xi_i, \pi_i)$ as in Equation (4). We again marginalize per-individual skipped cycles $s_{i,c_{new}}$ in

$$p(x_{i,c_{new}}|\lambda_i, \xi_i, \pi_i) = \sum_{s_{i,c_{new}}=0}^{s_{max}} p(x_{i,c_{new}}|s_{i,c_{new}}, \lambda_i, \xi_i)p(s_{i,c_{new}}|\pi_i). \quad (6)$$

One can readily compute the above via Monte Carlo, by drawing from the parameter posterior $p(\lambda_i, \xi_i, \pi_i|X_i, \Theta)$ as described in Equation (17) in Section A.3 of the Appendix, or via Importance Sampling by drawing from the prior $p(\lambda_i, \xi_i, \pi_i|\Theta)$ and weighting samples with $p(X_i|\lambda_i, \xi_i, \pi_i)$ as in Equation (4). After the training procedure, the hyperparameters Θ above will be replaced with the learned population level hyperparameters $\hat{\Theta}$.

In addition, we note that the above cycle length predictive posterior, as well as the skipping probability predictive posterior, can be updated as subsequent days of the next cycle pass by, which we have derived in Sections A.4 and A.5 of the Appendix, respectively.

4. Cohort

Real-world Menstrual mHealth Dataset⁴. We leverage a de-identified self-tracked dataset from Clue by BioWink (Clue, 2021), comprised of 117,014,597 self-tracking events over 378,694 users. Clue app users input overall personal information at sign-up, such as age and birth control type. The dataset contains information from 2015-2018 for users worldwide, covering countries within North and South America, Europe, Asia and Africa. In the entire dataset, the median number of tracked menstrual cycles is 11. Inclusion criteria into the cohort were: (1) users likely to have ovulatory cycles, that is aged 21-33 with natural cycle (i.e., no contraception); (2) users with at least 11 cycles tracked. The cohort resulted in 186,106 menstruators. For the experiments described in this paper, we randomly select a subset of 50,000 users. The summary statistics of the overall and the selected cohort for the experiments are provided in Table 1. We observe minimal differences in cycle length and period length statistics between the full cohort and the selected cohort.

4. Researchers interested in gaining access to the data can contact Clue by BioWink GmbH and establish a data use agreement with them.

Table 1: Summary statistics for the overall cohort and the 50,000 random user subset. Note that race/ethnicity information is not available from this de-identified dataset.

| Summary statistic | Full cohort | Selected cohort |
|---|---------------|-----------------|
| Total number of users | 186,106 | 50,000 |
| Total number of cycles | 2,047,166 | 550,000 |
| Cycle length in days: mean±sd (median) | 30.7±7.9 (29) | 30.6±7.7 (29) |
| Period length in days: mean±sd (median) | 4.1±1.8 (4) | 4.1±1.8 (4) |
| Age in years: mean±sd (median) | 25.6±3.6 (25) | 25.6±3.6 (25) |

Data Extraction and Feature choices. Even though Clue’s mHealth app users can self-track multiple symptoms over time, we focus on period data only, i.e., users’ self-reports on which days they have their period. A period is defined as sequential days of bleeding (greater than spotting and within ten days after the first greater than spotting bleeding event) unbroken by no more than one day on which only spotting or no bleeding occurred. We use cycle lengths as input to our proposed model, where we define a menstrual cycle as the span of days from the first day of a period through to and including the day before the first day of the next period (Vitzthum, 2009). We discard any cycle a user has indicated to be excluded from their history —e.g., if the user felt that the cycle was not representative of their typical menstruation due to a medical procedure or changes in birth control.

Synthetic Datasets⁵. To assess the ability of our model to recover ground truth (only possible with simulated data), we leverage two alternative generative processes. A Poisson generative model of cycle lengths, where the observed cycle lengths are drawn from the generative model by Li et al. (2021); i.e., cycle length data follows a Poisson distribution (see model and parameterization details in Appendix B.1). A Generalized Poisson generative model of cycle lengths, where the observed data are drawn from the generative model as proposed in Section 3; i.e., cycle length data is drawn from a Generalized Poisson distribution (full details are provided in Appendix B.2). For each of the simulated scenarios, we draw cycle length data for 50,000 users, with $C_i = 11$ cycles for each user.

5. Evaluation

5.1. Evaluation Approach: Real and Synthetic Study Designs

Our objective is to accurately predict the next cycle length of a mHealth menstrual app user, based on their previously-tracked cycle lengths. To that end, we train the proposed generative model (as described in Section 3) and several baselines (described in Section 5.1.1) on the cycle length information from each of the datasets described in Section 4.

We train on the first 10 cycle lengths of each user ($C_i = 10, \forall i$) and, given the hyperparameters $\hat{\Theta}$ learned via the training procedure, we predict the next cycle length (i.e., each user’s 11th cycle) via the predictive posterior in Section 3.2.2. Consequently, the train-test split is within, and not across, individuals: we train personalized models with each individual’s first 10 cycles, and evaluate our individualized predictions with respect to each user’s next cycle length.

5. The synthetic dataset can be generated with the Python codebase publicly available in https://github.com/iurteaga/menstrual_cycle_analysis.

We average our results (and provide standard deviations) over $k = 5$ realizations to aggregate over the inherent uncertainties of the training and testing procedure: e.g., robustness to random number generator seeds, Monte Carlo sampling and the optimization procedure.

We note that the learned predictive posterior $p(x_{i,c_{new}} = x | X_i, \hat{\Theta})$ provides per-user fully probabilistic predictions, i.e., it computes the probability of the next cycle length $x_{i,c_{new}}$ being of length $x \in \mathbb{N}$ for each user i . Therefore, we are enabled to provide both point estimate predictions (e.g., the mean or mode of the predictive distribution), as well as to evaluate how well-calibrated the predicted cycle length posterior is.

5.1.1. BASELINES

We compare the performance of our model to the following baselines⁶:

- **CNN**: a 2-layer convolutional neural network with a 3-dimensional kernel.
- **RNN**: a 2-layer bidirectional recurrent neural network with a 3-dimensional hidden state.
- **LSTM**: a 2-layer Long Short-Term Memory neural network with a 3-dimensional hidden state.
- **Poisson model**: the Poisson-based predictive model proposed by [Li et al. \(2021\)](#).

5.1.2. PREDICTION METRICS

We use several accuracy metrics for the evaluation of next cycle length point estimates $\widehat{x_{i,c_{new}}}$ with respect to true cycle lengths $x_{i,c_{new}}$ for all I users in the cohort. The root-mean squared error, $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^I (x_{i,c_{new}} - \widehat{x_{i,c_{new}}})^2}{I}}$; the median squared error, $\text{MedianSE} = \text{Median} \left[(x_{i,c_{new}} - \widehat{x_{i,c_{new}}})^2 \right]$ (which is less sensitive to outliers than the RMSE); the mean absolute error, $\text{MAE} = \frac{\sum_{i=1}^I |x_{i,c_{new}} - \widehat{x_{i,c_{new}}}|}{I}$; and the median absolute error, $\text{MedianAE} = \text{Median} \left[|x_{i,c_{new}} - \widehat{x_{i,c_{new}}}| \right]$.

5.1.3. CALIBRATION METRICS

We leverage a diverse set of calibration metrics and scoring rules, both visual and numeric, to evaluate the uncertainty quantification of the generative models’ predictive posteriors. On one hand, we consider the following, most often visually presented, calibration metrics:

The probability integral transform (PIT), defined as the value that the cumulative density function (CDF) of a predictive model $F(\cdot)$ attains at the observation, i.e., $p_i = F(x_i)$, where $x_i \sim g(\cdot)$ is drawn from the true (yet unknown) generating mechanism, with CDF $G(\cdot)$. For continuous true $G(\cdot)$ and predictive $F(\cdot)$, p_i has a uniform distribution if the predictions are ideal, i.e., if $F(\cdot) = G(\cdot)$. PITs are most often reported as a histogram over the set of observed instances $x_i, \forall i$; and for the ideal case, the histogram of the PIT values is (asymptotically) uniform. The uniformity of the PIT is a necessary, but not sufficient, condition for the predictive distribution to be ideal —[Gneiting et al. \(2007\)](#) provides a detailed explanation of PIT’s limitations. Visual assessment of PIT histograms

6. Similar to what is reported in [Li et al. \(2021\)](#), we don’t observe any significant performance difference with other architectures that incorporate higher kernel or hidden state dimensionalities.

provides insights into the calibration deficiencies of a predictive posterior: hump-shaped histograms indicate over-dispersion (i.e., prediction intervals are too wide), while U-shaped histograms correspond to under-dispersion (i.e., too narrow predictive distributions). Note that triangle-shaped histograms indicate biased predictive distributions. Since its proposal by Rosenblatt (1952), many authors have extended and studied PIT’s advantages and disadvantages —see Section 3 in (Gneiting et al., 2007).

The marginal calibration plot (MCP), defined as the difference between the predictive CDF $F(\cdot)$ and the empirical CDF of the observed data $\hat{G}(\cdot)$ (Gneiting et al., 2007), i.e., $F(x_i) - \hat{G}(x_i)$, $\forall x_i$. The most straightforward approach is to visualize the above difference over all observed instances, towards assessing the marginal calibration of the predictive distribution. The marginal calibration is concerned with the closeness between the predictive outcomes (i.e., predictive distribution) and the actual, observed outcomes (i.e., the data). The interested reader can find in Gneiting et al. (2007) a rigorous study of how, under mild regularity conditions, marginal calibration is a necessary and sufficient condition for the asymptotic equality between the average predictive CDF and the empirical CDF of the observations. Visually, one expects minor fluctuations under the hypothesis of marginal calibration (i.e., the MCP is almost flat), while major excursions from the origin indicate a lack of marginal calibration.

On the other hand, to quantify with a single numerical score how closely a model’s predictive distribution matches each user’s observed cycle lengths, we consider several scoring rules. The selected scores presented below form a comprehensive set of summary measures of predictive performance, as they address calibration and sharpness simultaneously (Gneiting et al., 2007). We note that all these are proper scoring rules⁷ in general, and strictly proper⁸ under quite general conditions —a detailed, theoretically grounded review of the above and other scoring rules is provided in (Gneiting and Raftery, 2007). We define them here for predictive distributions on the natural line $x \in \mathbb{N}$, but they are scoring rules that can readily accommodate continuous variables.

The quadratic or Brier score, defined as $\text{BrierS}(p, x_i) = -\sum_{x=1}^{x_{max}} (\delta(x - x_i) - p(x))^2$.

The pseudo-spherical score, defined as $\text{PseudoS}(p, x_i) = \frac{p(x_i)^{\alpha-1}}{\|p\|_{\alpha}^{\alpha-1}}$, which reduces to the more common **Spherical score** when $\alpha = 2$, used in our results.

The logarithmic score, defined as the log-likelihood of the observation x_i under the predicted posterior $p(\cdot)$: $\text{LogS}(p, x_i) = \log p(x_i)$, which relates directly to the negative Shannon entropy and the commonly used log-likelihood metric. Interestingly, this score emerges as a limiting case of the pseudospherical score with $\alpha \rightarrow 1$ when it is suitably scaled.

The continuous ranked probability score (CRPS), defined in terms of the CDF $F(\cdot)$ of the predictive posterior, i.e., $\text{CRPS}(F, x_i) = -\int_{-\infty}^{\infty} (F(x) - \mathbb{1}[x_i])^2 dx$, and corresponds to the integral of the **Brier scores** for the associated binary probability forecasts at all real-valued thresholds —note that when dealing with integers in the natural line, the thresholds are countable, resulting in a sum over a finite number of bins. The motivation for the CRPS is to overcome several limitations regarding other metrics on continuous variables. If Lebesgue

7. $S(P, P) \geq S(P, Q)$, $\forall P, Q \in \mathcal{P}$.

8. $S(P, P) = S(P, Q)$, if and only if $P = Q$.

densities on the real line are used to predict discrete observations, then the **logarithmic score** encourages the placement of artificially high density ordinates on the target values, and no credit is given for assigning high probabilities to values near but not identical to the one materializing. As such, defining scoring rules in terms of predictive CDFs (instead of probability density functions) has been argued for by [Gneiting et al. \(2007\)](#).

5.2. Validation on the synthetic cycle length dataset

We first showcase the added flexibility of our proposed method by leveraging the synthetic datasets described in Section 4. Our synthetic data results (presented in Section B) demonstrate that the proposed model provides better uncertainty quantification capabilities than the alternative proposed by [Li et al. \(2021\)](#), both in terms of predictive accuracy and calibration metrics. Specifically, when the cycle length data is Poisson distributed, both models can accurately fit the data and provide well-calibrated predictions: all scoring rules are identical for both models, see Section B.1.

On the contrary, when the observed cycle length data is drawn from a Generalized Poisson that is under-dispersed, we observe that the **Proposed model** clearly outperforms the **Poisson model** both in terms of predictive accuracy and calibration metrics. Specifically, we note (see Figures in Section B.2) that the PIT of the **Poisson model** is hump-shaped, i.e., it is clearly over-dispersed, while the **Proposed model**'s PIT histogram is close to a uniform distribution. In addition, the MCP plot for the **Proposed model** hardly deviates from the origin, while the **Poisson model** showcases a calibration mismatch around $x_{c_{new}} = 20$. Overall, these results validate that a Generalized Poisson based model is able to more flexibly adjust to the uncertainty of observed cycle lengths and provide well-calibrated predictions.

5.3. Results on the real-world mHealth dataset

We now present results for all the considered models (generative and neural network based, as described in Section 5.1.1) in the real-world cycle length dataset presented in Section 4. We provide in Table 2 point estimate results for all the models at the next day of the last observed cycle length (i.e., day 0 of the next cycle), as per the metrics in Section 5.1.2.

First, we conclude that in our use case, black-box neural network architectures do not provide any prediction accuracy advantage, which aligns with results presented by [Li et al. \(2021\)](#). Results showcasing the calibration shortcomings of the studied neural network architectures are presented in Appendix B.3. Second, we notice that our model's predictive accuracy is as good as the Poisson-based alternative of [Li et al. \(2021\)](#) when the mode of each model's predictive posterior is used as the predicted cycle length point estimate.

Third, as both the **Poisson model** and our **Proposed model** provide a full predictive posterior density over the set of natural integers $x_{i,c_{new}} \in \mathbb{N}$, we compare their performance when considering the mean and the mode of the predictive posterior as point estimates. We observe a slightly better performance of our **Proposed model**, both in cohort-level average results as well as in their variability, specially so for the metrics most robust to outliers (i.e., **MedianSE**, **MAE** and **MedianAE**). This performance difference suggests that the mean and the mode do not coincide in each model's predictive distributions, which we hypothesize is explained by the dispersion of such densities, i.e., the shape and width of their posterior densities around the mode.

Table 2: Real-world dataset: Point estimate results for all models

| Model | RMSE | MedianSE | MAE | MedianAE |
|-----------------------|-----------------------|------------------------|-----------------------|-----------------------|
| CNN | 7.243 (± 0.000) | 11.089 (± 0.341) | 4.379 (± 0.016) | 3.330 (± 0.051) |
| LSTM | 6.730 (± 0.017) | 4.303 (± 0.515) | 3.626 (± 0.082) | 2.071 (± 0.123) |
| RNN | 6.820 (± 0.062) | 5.606 (± 0.783) | 3.846 (± 0.071) | 2.362 (± 0.166) |
| Poisson model (mode) | 6.856 (± 0.023) | 4.000 (± 0.000) | 3.451 (± 0.006) | 2.000 (± 0.000) |
| Proposed model (mode) | 6.790 (± 0.007) | 4.000 (± 0.000) | 3.459 (± 0.003) | 2.000 (± 0.000) |
| Poisson model (mean) | 6.690 (± 0.002) | 4.818 (± 0.292) | 3.639 (± 0.042) | 2.194 (± 0.066) |
| Proposed model (mean) | 6.691 (± 0.005) | 4.237 (± 0.064) | 3.592 (± 0.009) | 2.058 (± 0.015) |

To elucidate the added flexibility of the **Proposed model** to capture cycle length uncertainty, we first visualize in Figure 2 the scatter plot of per-user true cycle length average (x-axis) versus each model’s per-user expected cycle length (y-axis). Each dot in Figure 2 represents a user, colored by the cycle length variability of each user: i.e., lighter color for users with low cycle length variability, darker color for users with high cycle length variability.

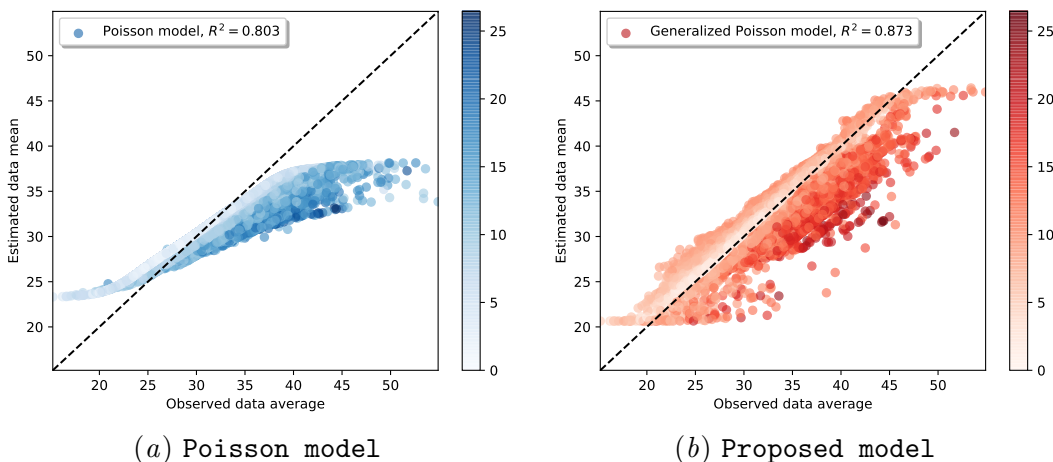


Figure 2: Real-world dataset: Fitting sufficient statistics of observed cycle length data, colored by user cycle length variability. The colorbar indicates the standard deviation of observed per-user cycle lengths.

We observe that the **Poisson model** struggles to find the right mean-variance balance: note how skewed the scatter plot in Figure 2(a)subfigure is, with most of the users (irrespective of their variability) situated below and to the right side of the $x = y$ line.

We hypothesize that this skewness is due to the rigid parameterization of a Poisson distribution: there is only one degree of freedom (i.e., λ_i) that determines both the mean and the variance of each user’s cycle lengths. On the contrary, the two-parameter Generalized Poisson can adjust, via λ_i and ξ_i , both the mean and the variance of per-user cycle lengths. As shown in Figure 2(b)subfigure, this ability to quantify the data uncertainty allows the **Proposed model** to fit the data better ($R^2 = 0.873$) than the **Poisson model** ($R^2 = 0.803$).

We now turn our attention to the full posterior predictive distribution of each of the models, to investigate their uncertainty quantification capabilities.

In Figure 3, we illustrate the cycle length predictive posterior of each generative model for a randomly selected user, as days of the subsequent cycle proceed (the form of this per-day cycle length posterior is described in Section A.4 of the Appendix). Note how the **Proposed model** consistently provides less uncertain (i.e., under-dispersed) predictions.

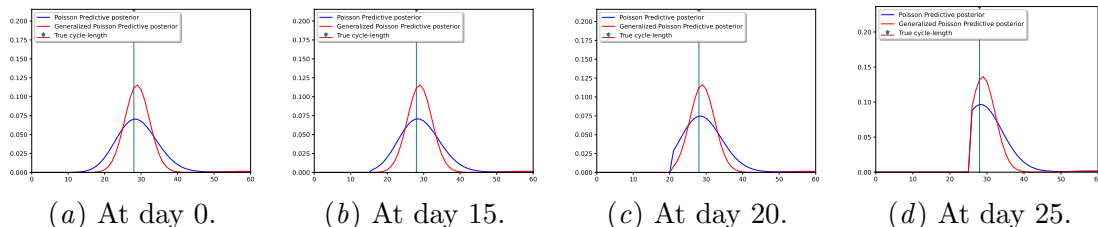


Figure 3: Real-world dataset: Predictive posteriors for a random user at different days of the next cycle.

Table 3: Real-world dataset: Posterior predictive width at day 0 of next cycle.

| Model | Posterior predictive interval width at level α . | | |
|----------------|---|-----------------------|------------------------|
| | $\alpha = 0.8$ | $\alpha = 0.5$ | $\alpha = 0.2$ |
| Poisson model | 2.925 (± 0.005) | 7.835 (± 0.026) | 15.248 (± 0.074) |
| Proposed model | 1.846 (± 0.014) | 4.940 (± 0.038) | 9.692 (± 0.089) |

We showcase in Table 3 that the under-dispersed predictive posterior of the **Proposed model** occurs for all users in the dataset, by providing the average posterior predictive width at level α . The values provided in the table indicate the width (in days) of the $(1-\alpha)$ centered probability mass, i.e., the width of the posterior predictive distribution between quantiles $\alpha/2$ and $1 - \alpha/2$, as illustrated in Figure 4, for the **Proposed model**'s posterior of Figure 3 with $\alpha = 0.5$ at day 0. Note that the posterior predictive width in Table 3 for the 20% posterior mass of our **Proposed model** is less than 2 days, while it's almost 3 for the **Poisson model**. Besides, the 50% posterior mass width of the **Poisson model** is almost 8 days (i.e., the next period is predicted to occur within an interval longer than a week), raising the question on how useful such prediction is for mHealth users.

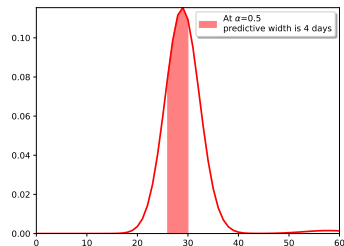


Figure 4. Posterior predictive width at day 0.

In order to settle our claim that the Generalized Poisson based model provides better calibrated predictions, we provide in Table 4 the average results for all the considered scoring rules described in Section 5.1.3, along with PIT and MCP plots in Figure 5: note how over-dispersed (hump-shaped) the **Poisson model** is in Figure 5(a) subfigure, and how, in Figure 5(b) subfigure, we observe a lack of posterior sharpness for the **Poisson model**. The calibration shortcomings of neural network based models are showcased in Appendix B.3.

Table 4: Real-world dataset: Calibration results for the generative models, higher is better.

| Model | Brier score | Spherical score | Logarithmic score | CRPS |
|----------------|------------------------|-----------------------|------------------------|------------------------|
| Poisson model | -0.931 (± 0.000) | 0.266 (± 0.000) | -3.022 (± 0.001) | -2.922 (± 0.003) |
| Proposed model | -0.910 (± 0.000) | 0.299 (± 0.000) | -2.855 (± 0.002) | -2.740 (± 0.001) |

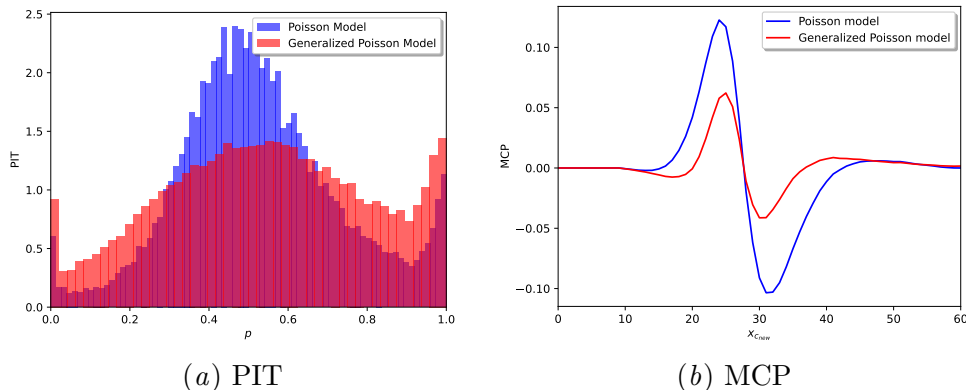


Figure 5: Real-world dataset: Calibration plots for a realization of each generative model.

As demonstrated across the variety of considered metrics, we can conclude that the **Proposed model** provides better calibrated results. However, as shown in Figure 5, the **Proposed model** is still not ideal: it does not result in a fully uniform PIT and its MCP slightly fluctuates away from 0 around the median cycle length (29 days) of the cohort.

To conclude, we emphasize that the **Proposed model** provides under-dispersed (i.e., less uncertain) and better calibrated cycle length predictions than the state-of-the-art alternative of Li et al. (2021), both as demonstrated at the user and population level. In addition, we showcase in Figure 6 how the under-dispersed predictive posterior of our **Proposed model** provides additional (across population and point estimate-based) predictive benefits: the predictive accuracy of the **Proposed model** is better than that of the **Poisson model** as days of the next cycle proceed. This is especially evident about a week before the median cycle length of the studied cohort. In other words, the **Proposed model** outperforms other models’ predictive accuracy 6 to 8 days before the next cycle length starts.

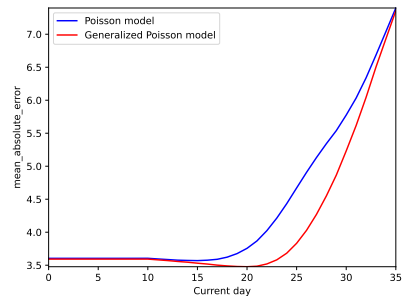


Figure 6. Real-world dataset: Prediction accuracy at different days of the next cycle.

6. Discussion

We proposed a flexible generative model that provides accurate, well-calibrated predictions of menstrual cycle lengths based on self-tracked mobile health data. Specifically, we investigated how to overcome certain limitations of a Poisson regression-based cycle length model by making use of a more flexible distribution, namely the Generalized Poisson. Our proposed model allows for accurate uncertainty quantification: it provides two degrees of freedom to fit the mean and variance of the observed cycle lengths. The model’s ξ parameter allows for controlling the dispersion of its predictive posterior, which enables better calibrated predictive posteriors, as demonstrated by our results.

Due to the well-calibrated predictions, the model not only yields improved predictive accuracy as the cycle days proceed (see Figure 6), but provides more meaningful (i.e., less uncertain) cycle length predictions (see Figure 3).

We argue that more certain cycle length predictions, like the ones provided by our proposed model, may benefit mHealth users: they help increase trust in the model and may mitigate user notification fatigue, a well-known phenomenon where users ignore prediction notifications if they occur too often and with low sensitivity.

More broadly, we argue that uncertainty quantification and calibration are critical in the domain of health and healthcare. One must account for the variability of the studied physiology, and adjust to the uncertainties of the data collection and reporting process. To that end, we have presented a diverse tool-set of calibration metrics that are of use in assessing the predictions of our proposed model, and argue that they should be readily incorporated into the practice of machine learning in healthcare.

We acknowledge several **limitations of our work**: (i) while we argue that less uncertain cycle length predictions may reduce mHealth user notification fatigue, we leave to future work to validate such a hypothesis; and (ii) our model includes features of the menstrual cycle only related to its length. While it is a minimal feature that we know will be present across many app users, there might be additional features like signs and symptoms of the menstrual cycle that may improve our predictive model.

Overall, our work showcases that generative models can accommodate the idiosyncrasies of mHealth data to provide well-calibrated, accurate predictions. Less uncertain cycle length predictions are beneficial for menstrual health researchers, mHealth users and developers.

Acknowledgments

The authors are deeply grateful to all Clue users whose de-identified data have been used for this study. Iñigo Urteaga and Noémie Elhadad are supported by NLM award R01 LM013043. Kathy Li is supported by NSF’s Graduate Research Fellowship Program Award #1644869. We also acknowledge computing resources from Columbia University’s Shared Research Computing Facility project, which is supported by NIH Research Facility Improvement Grant 1G20RR030893-01, and associated funds from the New York State Empire State Development, Division of Science Technology and Innovation (NYSTAR) Contract C090171, both awarded April 15, 2010.

References

- Ana Carolina Alba, Thomas Agoritsas, Michael Walsh, Steven Hanna, Alfonso Iorio, PJ Devereaux, Thomas McGinn, and Gordon Guyatt. Discrimination and calibration of clinical prediction models: users’ guides to the medical literature. *JAMA*, 318(14):1377–1384, 2017.
- Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H Shah, and Andrew Y Ng. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, pages 145–155. PMLR, 2020.
- Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176, 2016.

- I. Chen, S. Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. Probabilistic machine learning for healthcare. *ArXiv*, abs/2009.11087, 2020.
- Leonard Chiazze, Franklin T. Brayer, Jr. John J. Macisco, Margaret P. Parker, and Benefict J. Duffy. The Length and Variability of the Human Menstrual Cycle. *The Journal of the American Medical Association*, 203(6):377–380, 1968.
- Clue. Clue by BioWink GmbH, Adalbertstraße 7-8, 10999 Berlin, Germany. <https://helloclue.com/>, 2021.
- Prem C Consul. *Generalized Poisson distributions: properties and applications*. M. Dekker, 1989.
- Prem C Consul and Felix Famoye. Generalized poisson distribution. *Lagrangian Probability Distributions*, pages 165–190, 2006.
- A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.
- A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.
- Francis X Diebold and Robert S Mariano. Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144, 2002.
- Michael W. Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M. Dai. Analyzing the Role of Model Uncertainty for Electronic Health Records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 204–213, New York, NY, USA, 2020.
- Daniel A Epstein, Nicole B Lee, Jennifer H Kang, Elena Agapie, Jessica Schroeder, Laura R Pina, James Fogarty, Julie A Kientz, and Sean A Munson. Examining Menstrual Tracking to Inform the Design of Personal Informatics Tools. *Proceedings of the SIGCHI conference on human factors in computing systems. CHI Conference*, 2017:6876–6888, May 2017.
- Rebecca J Ferrell, Kathleen A O'Connor, German Rodriguez, Tristan Gorrindo, Darryl J Holman, Eleanor Brindle, Rebecca C Miller, et al. Monitoring reproductive aging in a 5-year prospective study: aggregate and individual changes in steroid hormones and menstrual cycle lengths with age. *Menopause*, 12:567–757, 2005.
- Andrew YK Foong, David R Burt, Yingzhen Li, and Richard E Turner. On the expressiveness of approximate inference in bayesian neural networks. *arXiv preprint arXiv:1909.00719*, 2019.
- Sarah Fox and Daniel A Epstein. Monitoring menses: Design-based investigations of menstrual tracking applications. *The Palgrave Handbook of Critical Menstruation Studies*, pages 733–750, 2020.
- R Gillon. Medical ethics: four principles plus attention to scope. *BMJ*, 309(6948):184, 1994.

- Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.
- Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.
- Mark Goldstein, Xintian Han, Aahlad Puli, Adler J Perotte, and Rajesh Ranganath. X-CAL: Explicit Calibration for Survival Analysis. *arXiv preprint arXiv:2101.05346*, 2021.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- Paul KJ Han, William MP Klein, and Neeraj K Arora. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838, 2011.
- Siobán D. Harlow, Margery Gass, Janet E. Hall, Roger Lobo, Pauline Maki, Robert W. Rebar, Sherry Sherman, Patrick M. Sluss, Tobie J.de Villiers, and for the STRAW+10 Collaborative Group. Executive Summary of the Stages of Reproductive Aging Workshop + 10: Addressing the Unfinished Agenda of Staging Reproductive Aging. *The Journal of Clinical Endocrinology & Metabolism*, 97(4):1159–1168, 04 2012.
- Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.
- Kathy Li, Iñigo Urteaga, Chris H. Wiggins, Anna Druet, Amanda Shea, Virginia J. Vitzthum, and Noémie Elhadad. Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile health data. *Nature Digital Medicine*, 3(79), 2020.
- Kathy Li, Iñigo Urteaga, Amanda Shea, Virginia J Vitzthum, Chris H Wiggins, and Noémie Elhadad. A generative, predictive model for menstrual cycle lengths that accounts for potential self-tracking artifacts in mobile health data. *arXiv preprint arXiv:2102.12439*, 2021. Presented at the Machine Learning for Mobile Health Workshop at NeurIPS 2020.
- Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.
- Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

- Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring Calibration in Deep Learning. In *CVPR Workshops*, volume 2, 2019.
- Rosemary Orchard. Apple’s Cycle Tracking: A Personal Review. <https://www.macstories.net/stories/apples-cycle-tracking-a-personal-review/>, November 2019.
- Emma Pierson, Tim Althoff, and Jure Leskovec. Modeling Individual Cyclic Variation in Human Behavior. In *Proceedings of the 2018 World Wide Web Conference, WWW ’18*, pages 107–116, Republic and Canton of Geneva, Switzerland, 2018.
- Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.
- Wendy A Rogers and Mary J Walker. Fragility, uncertainty, and healthcare. *Theoretical medicine and bioethics*, 37(1):71–83, 2016.
- Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.
- Uwe Siebert. When should decision-analytic modeling be used in the economic evaluation of health care?, 2003.
- Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019.
- Ilias Soumpasis, Bola Grace, and Sarah Johnson. Real-life insights on menstrual cycles and ovulation using big data. *Human Reproduction Open*, 2020(2), 04 2020.
- Richard J Stevens and Katrina K Poppe. Validation of clinical prediction models: what does the ”calibration slope” really measure? *Journal of clinical epidemiology*, 118:93–99, February 2020.
- Laura Symul, Katarzyna Wac, Paula Hillard, and Marcel Salathe. Assessment of Menstrual Health Status and Evolution through Mobile Apps for Fertility Awareness. *bioRxiv*, 2018.
- Alan E. Treloar, Ruth E. Boynton, Borghild G. Behn, and Byron W. Brown. Variation of the human menstrual cycle through reproductive life. *International journal of fertility*, 12(1 Pt 2):77–126, 1967.
- Iñigo Urteaga, Mollie McKillop, and Noémie Elhadad. Learning endometriosis phenotypes from patient-generated data. *npj Digital Medicine*, 3(88), 06 2020.
- Werner Vach. Calibration of clinical prediction rules does not just assess bias. *Journal of clinical epidemiology*, 66(11):1296–1301, 2013.
- Virginia J. Vitzthum. The ecology and evolutionary endocrinology of reproduction in the human female. *American Journal of Physical Anthropology*, 140(S49):95–136, 2009.

Ellen Ann Wartella, Vicky Rideout, Heather Montague, Leanne Beaudoin-Ryan, and Alexis Re Lauricella. Teens, health and technology: A national survey. *Media and Communication*, 4(3):13–23, 1 2016.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Andrew Gordon Wilson. The case for Bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for Bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.

Appendix A. Methods

A.1. The Generalized Poisson

The Generalized Poisson is a distribution that belongs to the class of Lagrangian distributions over discrete, non-negative integers, with parameters λ and ξ that are independent (Consul, 1989; Consul and Famoye, 2006).

The probability density function (pdf) of a Generalized Poisson, hereby denoted as $\mathcal{GP}(\cdot)$, follows

$$\mathcal{GP}(x|\lambda, \xi) = \begin{cases} \lambda(\lambda + \xi x)^{x-1} \frac{e^{-\lambda - \xi x}}{x!}, & \lambda > 0, x = 0, 1, \dots, \\ 0, & \text{for } x > x_{max} \text{ if } \xi < 0, \end{cases} \quad (7)$$

where limits on λ and ξ are imposed to ensure that there are at least five classes with nonzero probability (Consul, 1989).

The first two moments of a $\mathcal{GP}(x|\lambda, \xi)$ follow

$$\mu_x = \frac{\lambda}{(1 - \xi)}, \quad (8a)$$

$$\sigma_x^2 = \frac{\lambda}{(1 - \xi)^3}, \quad (8b)$$

and other moments of interest can be computed in closed form, see (Consul and Famoye, 2006) for a full characterization of this distribution.

The Generalized Poisson can be over- or under-dispersed, depending on the value of ξ : when $\xi = 0$, we recover the Poisson distribution. Specifically, for $\xi < 0$, the Generalized Poisson is under-dispersed (in comparison to a Poisson distribution) and it can be truncated to a maximum value x_{max} of x , requiring an additional normalizing factor $Z_{\mathcal{GP}(\lambda, \xi, x_{max})}$:

$$x \sim \mathcal{GP}(x|\lambda, \xi) = \begin{cases} \frac{\lambda(\lambda + \xi x)^{x-1} \frac{e^{-\lambda - \xi x}}{x!}}{Z_{\mathcal{GP}(\lambda, \xi, x_{max})}}, & x = 0, 1, \dots, \\ 0, & \text{for } x > x_{max} \text{ if } \xi < 0, \end{cases} \quad (9)$$

with

$$Z_{\mathcal{GP}(\lambda, \xi, x_{max})} = \sum_{x=0}^{x_{max}} \lambda(\lambda + \xi x)^{x-1} \frac{e^{-\lambda - \xi x}}{x!}. \quad (10)$$

A.2. The hierarchical, generative process for observed cycle lengths

The proposed model, depicted in Figure 1, is a generative process with the following random (observed and unobserved) variables and parameters:

- The observed variables are the cycle lengths x_{i,c_i} , with $c_i = \{1, \dots, C_i\}$ cycle lengths for each individual $i = \{1, \dots, I\}$.
- We denote with s_{i,c_i} the (latent) number of skipped (not-reported) cycles, with $c_i = \{1, \dots, C_i\}$ cycle lengths for each individual $i = \{1, \dots, I\}$.
- λ_i and ξ_i denote the Generalized Poisson parameters for each individual $i = \{1, \dots, I\}$.

- π_i are the parameters defining the per-individual $i = \{1, \dots, I\}$ probability of skipping a cycle.
- κ, γ are the population-level hyperparameters of a Gamma distribution prior over the λ_i .
- α_ξ, β_ξ are the population-level hyperparameters of a Beta distribution prior over the ξ_i .
- α, β are the population-level hyperparameters of a Beta distribution prior over the skipping probabilities π_i .

We now describe in detail the generative process of the proposed probabilistic model. First, one draws individual cycle length and self-tracking probability parameters from the population level distributions:

1. The parameter λ_i of each individual's Generalized Poisson is drawn from a population-level Gamma distribution with hyperparameters κ and γ

$$\lambda_i \sim p(\lambda|\kappa, \gamma) = \mathcal{G}(\lambda|\kappa, \gamma) = \frac{\gamma^\kappa}{\Gamma(\kappa)} \lambda^{\kappa-1} e^{-\gamma\lambda}, \quad \text{for } \lambda > 0 \text{ and } \kappa, \gamma > 0. \quad (11)$$

2. The parameter ξ_i of each individual's Generalized Poisson is drawn, conditioned on each λ_i , from a shifted and scaled population-level Beta distribution with hyperparameters α_ξ and β_ξ , so that $\xi_i \in \left[\max\{-1, \frac{-\lambda_i}{x_{max}}\}, 1 \right]$:

$$\xi_i \sim \max\left\{-1, \frac{-\lambda_i}{x_{max}}\right\} + \left(1 - \max\left\{-1, \frac{-\lambda_i}{x_{max}}\right\}\right) \mathcal{B}(\xi|\alpha_\xi, \beta_\xi), \quad \text{for } \xi \in [0, 1], \quad (12)$$

$$\text{where } \mathcal{B}(\xi|\alpha_\xi, \beta_\xi) = \frac{\Gamma(\alpha_\xi + \beta_\xi)}{\Gamma(\alpha_\xi)\Gamma(\beta_\xi)} \xi^{\alpha_\xi-1} (1-\xi)^{\beta_\xi-1}, \quad \text{for } \xi \in [0, 1] \text{ and } \alpha_\xi, \beta_\xi > 0.$$

3. The probability π_i of each individual forgetting to track a period is drawn from a population-level Beta distribution with hyperparameters α and β ,

$$\pi_i \sim p(\pi|\alpha, \beta) = \mathcal{B}(\pi|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1}, \quad \text{for } \pi \in [0, 1] \text{ and } \alpha, \beta > 0. \quad (13)$$

Given per-individual parameters λ_i, ξ_i, π_i , then:

4. The number of cycles a user forgets to track s_{i,c_i} is drawn from a Truncated Geometric distribution with parameter π_i , i.e.,

$$s_{i,c_i} \sim p(s|\pi_i) = \frac{\pi_i^s (1 - \pi_i)}{\sum_{s=0}^{s_{max}} \pi_i^s (1 - \pi_i)} = \frac{\pi_i^s}{\sum_{s=0}^{s_{max}} \pi_i^s} = \frac{\pi_i^s (1 - \pi_i)}{(1 - \pi_i^{(s_{max}+1)})}, \quad \text{for } s \in \mathbb{N} \quad (14)$$

5. Each true (unobserved) cycle length x is drawn from a Generalized Poisson distribution parameterized with per-individual λ_i and ξ_i , i.e.,

$$x \sim p(x|\lambda_i, \xi_i) = \mathcal{GP}(x|\lambda_i, \xi_i) = \begin{cases} \lambda_i(\lambda_i + \xi_i x)^{x-1} \frac{e^{-\lambda_i - \xi_i x}}{x!}, & x = 0, 1, \dots, \\ 0, & \text{for } x > x_{max} \text{ if } \xi_i < 0. \end{cases} \quad (15)$$

6. Finally, the observed cycle length x_{i,c_i} for user i is drawn from a Generalized Poisson distribution, conditioned on the number of skipped cycles s_{i,c_i} , i.e.,

$$x_{i,c_i} \sim p(x|s_{i,c_i}, \lambda_i, \xi_i) = \sum_{s=0}^{s_{i,c_i}} p(x|s, \lambda_i, \xi_i) = \mathcal{GP}(x|(s_{i,c_i} + 1)\lambda_i, \xi_i) . \quad (16)$$

Note that this distribution results from the property of Generalized Poissons that the sum of two independent Generalized Poisson variables $X \sim \mathcal{GP}(\lambda_1, \xi)$ and $Y \sim \mathcal{GP}(\lambda_2, \xi)$ follows another Generalized Poisson: $X + Y \sim \mathcal{GP}(\lambda_1 + \lambda_2, \xi)$, ([Consul and Famoye, 2006](#)).

A.3. The model's posterior parameter distributions

In a similar approach to the joint data likelihood, we can analytically marginalize the skipped cycles and compute a MC approximation to each parameter posterior, as described below:

$$p(\lambda_i, \xi_i, \pi_i | X_i, \Theta) = \frac{p(X_i, \lambda_i, \xi_i, \pi_i | \Theta)}{p(X_i | \Theta)} = \frac{p(X_i, \lambda_i, \xi_i, \pi_i | \Theta)}{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} p(X_i, \lambda_i, \xi_i, \pi_i | \Theta) d\lambda_i d\xi_i d\pi_i} \quad (17a)$$

$$= \frac{p(X_i | \lambda_i, \xi_i, \pi_i) p(\lambda_i, \xi_i, \pi_i | \Theta)}{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} p(X_i | \lambda_i, \xi_i, \pi_i) p(\lambda_i, \xi_i, \pi_i | \Theta) d\lambda_i d\xi_i d\pi_i} \quad (17b)$$

$$= \frac{p(X_i | \lambda_i, \xi_i, \pi_i) p(\lambda_i | \kappa, \gamma) p(\xi_i | \lambda_i, \alpha_\xi, \beta_\xi) p(\pi_i | \alpha, \beta)}{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} p(X_i | \lambda_i, \xi_i, \pi_i) p(\lambda_i | \kappa, \gamma) p(\xi_i | \lambda_i, \alpha_\xi, \beta_\xi) p(\pi_i | \alpha, \beta) d\lambda_i d\xi_i d\pi_i} \quad (17c)$$

$$\approx \sum_{m_\lambda=1}^{M_\lambda} \sum_{m_\xi=1}^{M_\xi} \sum_{m_\pi=1}^{M_\pi} \frac{p(X_i | \lambda_i^{(m_\lambda)}, \xi_i^{(m_\xi)}, \pi_i^{(m_\pi)})}{\sum_{m_\lambda=1}^{M_\lambda} \sum_{m_\xi=1}^{M_\xi} \sum_{m_\pi=1}^{M_\pi} p(X_i | \lambda_i^{(m_\lambda)}, \xi_i^{(m_\xi)}, \pi_i^{(m_\pi)})} , \quad (17d)$$

$$\text{with } p(X_i | \lambda_i^{(m_\lambda)}, \xi_i^{(m_\xi)}, \pi_i^{(m_\pi)}) = \prod_{c_i=1}^{C_i} p(x_{i,c_i} | \lambda_i^{(m)}, \xi_i^{(m)}, \pi_i^{(m)}) \quad (17e)$$

$$= \prod_{c_i=1}^{C_i} \sum_{s_{i,c_i}=0}^{s_{max}} p(x_{i,c_i} | s_{i,c_i}, \lambda_i^{(m)}, \xi_i^{(m)}) p(s_{i,c_i} | \pi_i^{(m)}) , \quad (17f)$$

$$\text{and } \begin{cases} \lambda_i^{(m_\lambda)} \sim p(\lambda_i | \kappa, \gamma) , m_\lambda = 1, \dots, M_\lambda , \\ \xi_i^{(m_\xi)} \sim p(\xi_i | \lambda_i^{(m_\lambda)} \alpha_\xi, \beta_\xi) , m_\xi = 1, \dots, M_\pi , \\ \pi_i^{(m_\pi)} \sim p(\pi_i | \alpha_\pi, \beta_\pi) , m_\pi = 1, \dots, M_\pi . \end{cases} \quad (17g)$$

A.4. The model’s predictive cycle length posterior by day

Our model allows for updating next cycle length predictions as each day of the next cycle passes. To that end, we compute the cycle length predictive posterior conditioned on x , the day of the cycle the user is currently on:

$$p(x_{i,c_{new}} | x_{i,c_{new}} > x, X_i, \Theta) = \frac{p(x_{i,c_{new}}, x_{i,c_{new}} > x | X_i, \Theta)}{p(x_{i,c_{new}} > x | X_i, \Theta)} \quad (18a)$$

$$= \frac{p(x_{i,c_{new}} | X_i, \Theta) I(x_{i,c_{new}} > x)}{p(x_{i,c_{new}} > x | X_i, \Theta)} \quad (18b)$$

since $p(x_{i,c_{new}}, x_{i,c_{new}} > x | X_i, \Theta) = 0$ if $x_{i,c_{new}} \leq x$. Note that the key term above is $p(x_{i,c_{new}} | X_i, \Theta)$, which follows the expression in (5).

A.5. The model’s predictive skipping probability posterior by day

Our model allows for updating per individual next cycle’s skipping probability predictions as each day of the next cycle passes. To that end, we compute the predictive posterior of skipping probabilities conditioned on x , the day of the cycle the user is currently on:

$$p(s_{i,c_{new}} | x_{i,c_{new}} > x, X_i, \Theta) = \frac{p(s_{i,c_{new}}, x_{i,c_{new}} > x | X_i, \Theta)}{p(x_{i,c_{new}} > x | X_i, \Theta)} \quad (19a)$$

$$= \frac{p(x_{i,c_{new}} > x | s_{i,c_{new}}, X_i, \Theta) p(s_{i,c_{new}} | X_i, \Theta)}{\sum_{s_{i,c_{new}}} p(x_{i,c_{new}} > x | s_{i,c_{new}}, X_i, \Theta) p(s_{i,c_{new}} | X_i, \Theta)} \quad (19b)$$

$$= \frac{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} p(x_{i,c_{new}} > x | s_{i,c_{new}}, \lambda_i, \xi_i) p(s_{i,c_{new}} | \pi_i) p(\lambda_i, \xi_i, \pi_i | X_i, \Theta) d\lambda_i d\xi_i d\pi_i}{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} \sum_{s_{i,c_{new}}} p(x_{i,c_{new}} > x | s_{i,c_{new}}, \lambda_i, \xi_i) p(s_{i,c_{new}} | \pi_i) p(\lambda_i, \xi_i, \pi_i | X_i, \Theta) d\lambda_i d\xi_i d\pi_i} \quad (19c)$$

$$= \frac{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} \sum_x p(x_{i,c_{new}} = x | s_{i,c_{new}}, \lambda_i, \xi_i) p(s_{i,c_{new}} | \pi_i) p(\lambda_i, \xi_i, \pi_i | X_i, \Theta) d\lambda_i d\xi_i d\pi_i}{\int_{\lambda_i} \int_{\xi_i} \int_{\pi_i} \sum_{s_{i,c_{new}}} \sum_x p(x_{i,c_{new}} = x | s_{i,c_{new}}, \lambda_i, \xi_i) p(s_{i,c_{new}} | \pi_i) p(\lambda_i, \xi_i, \pi_i | X_i, \Theta) d\lambda_i d\xi_i d\pi_i} \quad (19d)$$

where $p(s_{i,c_{new}} | \pi_i) = \frac{\pi_i^{s_{i,c_{new}}} (1-\pi_i)^{s_i - s_{i,c_{new}}}}{\sum_{s'} \pi_i^{s'} (1-\pi_i)^{s_i - s'}}$ and $p(\lambda_i, \xi_i, \pi_i | X_i, \Theta)$ is the parameter posterior after observing per-individual data X_i . We can compute the above via Monte Carlo —by drawing from the parameter posterior $p(\lambda_i, \xi_i, \pi_i | X_i, \Theta)$ as in Equation (17)— or via Importance Sampling —by drawing from the prior $p(\lambda_i, \xi_i, \pi_i | \Theta)$ and weighting them with $p(X_i | \lambda_i, \xi_i, \pi_i)$ as in Equation (4).

A.6. The model’s computational complexity and distributed training

The complexity of the training procedure of the proposed Generalized Poisson-based model is determined by the type-II maximum likelihood estimation of model hyperparameters. Specifically, it requires (i) the computation of the marginalized joint data likelihood in Equation (3), and (ii) finding the hyperparameters Θ that maximize Equation (3).

To that end, we leverage Monte Carlo integration and automatic differentiation. The fitting procedure is implemented and executed using stochastic gradient descent methods (as provided by most modern software packages). This same procedure is used for fitting the Poisson and Generalized Poisson-based alternatives.

For the underdispersed Generalized Poisson model (when $\xi < 0$), one must numerically compute the normalizing constant in Equation (10), which adds computational complexity. As a result, the per-epoch computational cost depends on the number of Monte Carlo samples used (M) in Equation (3), the accuracy on marginalizing out the skipped cycle probability (s_{max}) in Equation (4), and the computation of the per-user and parameter normalizing constant in Equation (10), dependent on the maximum cycle length (x_{max}).

We note that the hierarchical nature of the proposed model provides distributed learning opportunities: mHealth users do not need to share their data (they can locally compute their individualized predictions), and only need to share per-user data log-likelihood estimates for population-level hyperparameter inference.

When finding population-level hyperparameters, each user must only share its marginalized data likelihood in Equation (4), averaged over local parameter Monte Carlo samples, for which each user only needs access to the population-level hyperparameters Θ .

This training process can be executed in a distributed and iterative fashion, separating (on-device) per-user computations from global (centralized) hyperparameter searches, with not only computational, but privacy benefits too.

Appendix B. Results

We showcase the added flexibility of our proposed method, by leveraging the datasets described in Section 4 and detailed here.

Our goal is to demonstrate that the proposed model provides, both in synthetic and real datasets, better uncertainty quantification capabilities than the alternative proposed in Li et al. (2021) and other neural network based baselines.

B.1. Poisson generative model of cycle lengths

Data generating process. The observed cycle lengths are drawn from the generative model as in (Li et al., 2021), where cycle length data is assumed to obey a Poisson distribution. The specific hyperparameters used are $\Theta = \{\kappa = 180, \gamma = 6, \alpha = 2, \beta = 20\}$, resulting in parameter priors and per-individual sample draws as illustrated in Figure 7.

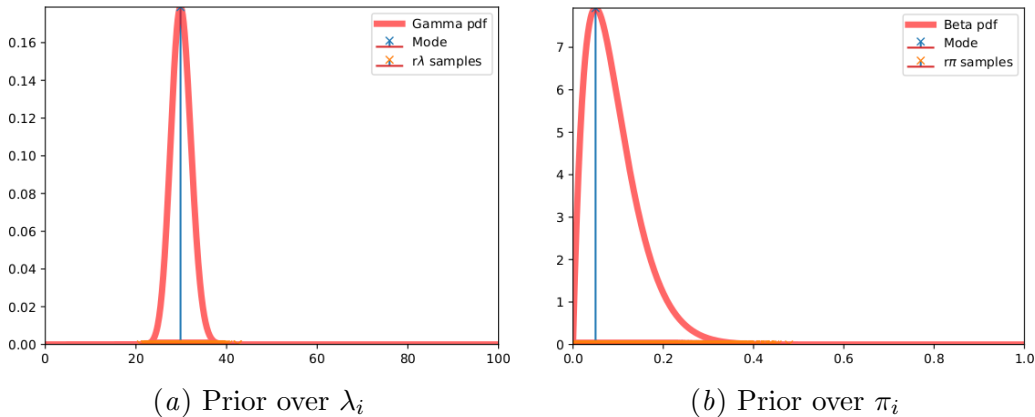


Figure 7: Synthetic Poisson: Ground truth parameter priors and per-individual drawn samples.

Predictive accuracy and calibration. Figures 8 and 9 demonstrate how our Proposed model is equivalent to the alternative Poisson model, when the data generating mechanism is indeed Poisson. The PIT of both models is almost uniform, and the MCP plots hardly deviate from the origin. Besides, note how the interval width for both models is identical in Table 6. This behavior demonstrates that, when the cycle length data is indeed Poisson, both models can accurately fit the data and provide well-calibrated predictions—all scoring rules in Table 5 are identical for both models.

Table 5: Synthetic Poisson: Calibration results for the generative models, higher is better

| Model | Brier score | Spherical score | Logarithmic score | CRPS |
|----------------|------------------------|-----------------------|------------------------|------------------------|
| Poisson model | -0.958 (± 0.000) | 0.204 (± 0.000) | -3.482 (± 0.000) | -5.381 (± 0.000) |
| Proposed model | -0.958 (± 0.000) | 0.204 (± 0.000) | -3.482 (± 0.000) | -5.382 (± 0.001) |

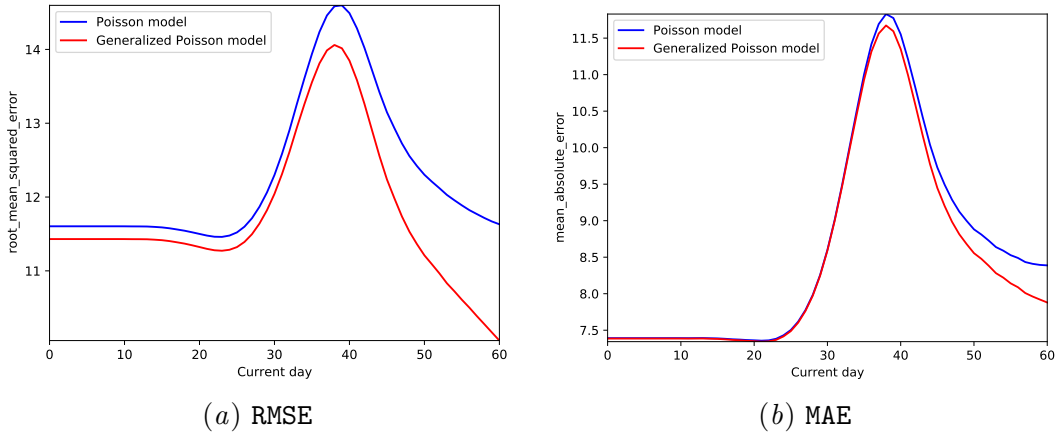


Figure 8: Synthetic Poisson: Prediction accuracy of the generative models at different days of the next cycle.

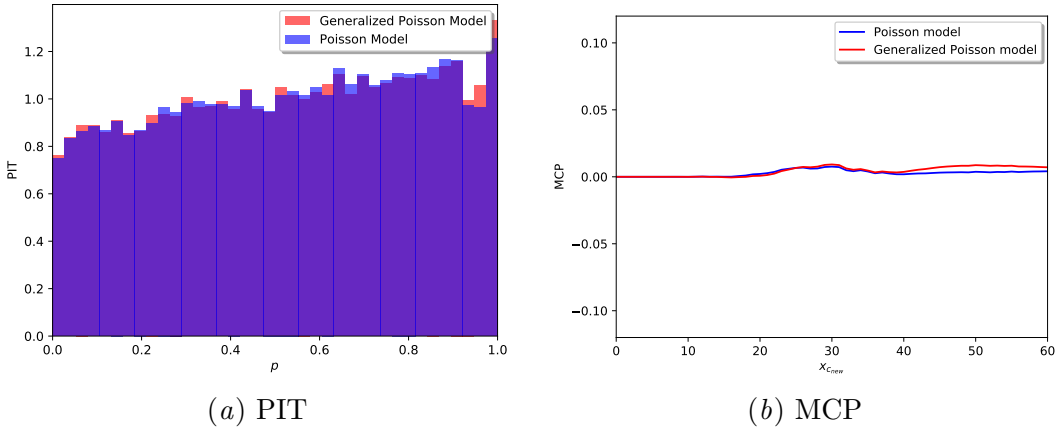


Figure 9: Synthetic Poisson: Calibration plots for a realization of each model.

Table 6: Posterior predictive width at day 0 of next cycle.

| Model | Interval Width for $(1 - \alpha)$ posterior mass. | | |
|----------------|---|-----------------------|------------------------|
| | $(1 - \alpha) = 0.2$ | $(1 - \alpha) = 0.5$ | $(1 - \alpha) = 0.8$ |
| Poisson model | 3.208 (± 0.006) | 8.757 (± 0.027) | 21.060 (± 0.167) |
| Proposed model | 3.211 (± 0.010) | 8.786 (± 0.030) | 21.182 (± 0.392) |

B.2. Generalized Poisson generative model of cycle lengths

Data generating process. The observed cycle lengths are drawn from the generative model as proposed in Section 3, where cycle length data is drawn from a Generalized Poisson distribution. The specific hyperparameters used for our simulation are $\Theta = \{\kappa = 160, \gamma = 4, \alpha_\xi = 2, \beta_\xi = 20, \alpha = 2, \beta = 20\}$, resulting in parameter priors and per-individual sample draws as illustrated in Figure 10.

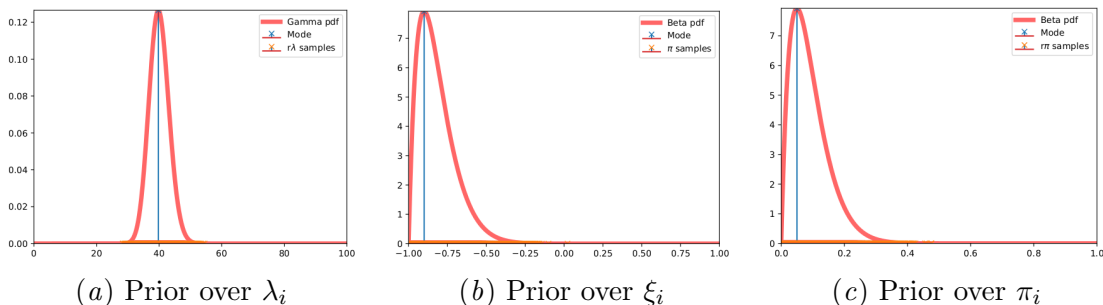


Figure 10: Synthetic Generalized Poisson: Ground truth parameter priors and per-individual drawn samples.

Predictive accuracy and calibration. Figures 11 and 12 demonstrate the limitations of the `Poisson model` when the data generating mechanism is not Poisson distributed: the generated cycle length data in these experiments is drawn from a Generalized Poisson that is under-dispersed (see specific hyperparameters above).

We observe that the `Proposed model` clearly outperforms the `Poisson model` both in terms of predictive accuracy (reduced MAE and RMSE in Figure 11) and calibration metrics (Figure 12 and Table 7).

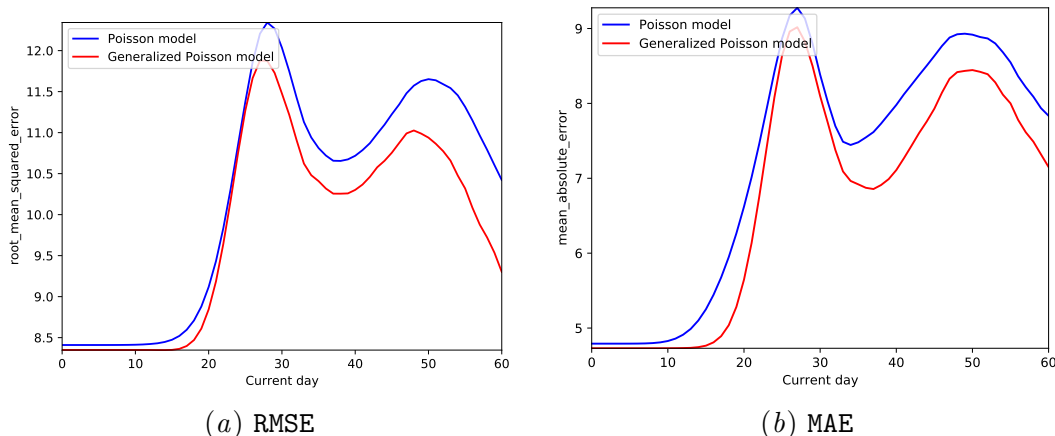


Figure 11: Synthetic Generalized Poisson: Prediction accuracy of the generative models at different days of the next cycle.

In Figure 12, note how the PIT of the `Poisson model` is hump-shaped, i.e., it is clearly over-dispersed, while the `Proposed model`'s PIT histogram is close to a uniform distribution. In addition, the MCP plot for the `Proposed model` hardly deviates from the origin, while the `Poisson model` showcases a calibration mismatch around $x_{c_{new}} = 20$.

Overall, these results validate our claim that a Generalized Poisson based model is able to more flexibly adjust to the uncertainty of observed cycle lengths and provide well-calibrated predictions—all scoring rule results in Table 5 are better for the `Proposed model`, and their posterior predictive interval width much smaller, see Table 8.

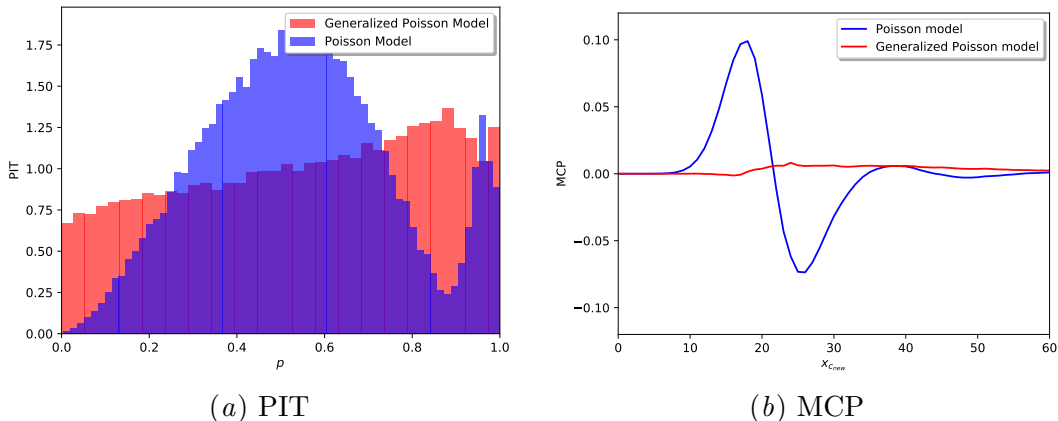


Figure 12: Synthetic Generalized Poisson: Calibration plots for a realization of each model.

Table 7: Synthetic Generalized Poisson: Calibration results for the generative models, higher is better

| Model | Brier score | Spherical score | Logarithmic score | CRPS |
|----------------|------------------------|-----------------------|------------------------|------------------------|
| Poisson model | -0.930 (± 0.000) | 0.269 (± 0.000) | -2.835 (± 0.001) | -3.553 (± 0.003) |
| Proposed model | -0.912 (± 0.000) | 0.295 (± 0.000) | -2.785 (± 0.001) | -3.385 (± 0.002) |

Table 8: Posterior predictive width at day 0 of next cycle.

| Model | Interval Width for $(1 - \alpha)$ posterior mass. | | |
|----------------|---|-----------------------|------------------------|
| | $(1 - \alpha) = 0.2$ | $(1 - \alpha) = 0.5$ | $(1 - \alpha) = 0.8$ |
| Poisson model | 2.761 (± 0.010) | 7.558 (± 0.034) | 17.569 (± 0.157) |
| Proposed model | 1.511 (± 0.005) | 4.141 (± 0.019) | 12.379 (± 0.199) |

B.3. Real-world Menstrual mHealth Data

We present predictive results for the models as described in Section 5.1.1 in the real-world cycle length dataset presented in Section 4.

Predictive accuracy and calibration. Since we have provided evidence in the main manuscript (see Table 2) on the point estimate accuracy of the neural network based alternatives, we hereby focus on their calibration limitations.

Table 9: Real-world dataset: Calibration results for the studied models, higher is better

| Model | Brier score | Spherical score | Logarithmic score | CRPS |
|----------------|------------------------|-----------------------|------------------------|------------------------|
| CNN | -1.816 (± 0.000) | 0.092 (± 0.000) | N/A | -4.274 (± 0.000) |
| LSTM | -1.790 (± 0.036) | 0.105 (± 0.018) | N/A | -3.867 (± 0.168) |
| RNN | -1.786 (± 0.028) | 0.107 (± 0.014) | N/A | -3.914 (± 0.179) |
| Poisson model | -0.931 (± 0.000) | 0.267 (± 0.000) | -3.022 (± 0.000) | -2.921 (± 0.001) |
| Proposed model | -0.910 (± 0.000) | 0.298 (± 0.000) | -2.854 (± 0.000) | -2.740 (± 0.001) |

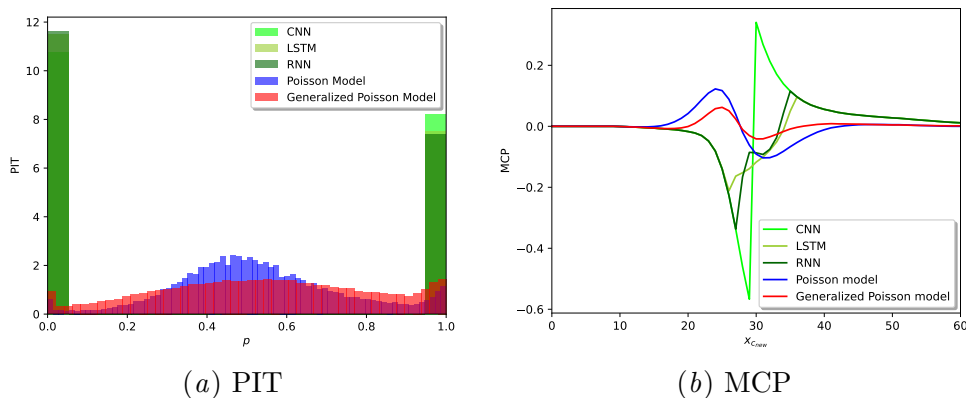


Figure 13: Real-world dataset: Calibration plots for a realization of each model.

Results in Table 9 and Figure 13 show very poor calibration performance of all the neural network based approaches, in alignment with the existing literature on the calibration limitations of these techniques in other applications as well (Guo et al., 2017; Nixon et al., 2019). Note that these results are computed based on the deterministic outputs of these models, i.e., the outputs of the CNN, LSTM and RNN models are point estimates, hence the extreme-valued PIT results in Figure 13(a) subfigure.

We acknowledge that this behavior could be avoided with Bayesian or ensemble-based neural network models that provide probabilistic outputs. However, implementing those alternative models was out of the scope of this work, and we reiterate that there is a growing literature on the calibration shortcomings of these approaches (Wenzel et al., 2020), specially so when approximate inference is used (Foong et al., 2019).

As demonstrated across the variety of considered metrics, we conclude that the **Proposed model** provides better calibrated results.

Computational complexity. As explained in Section A.6, the complexity of the training procedure of the proposed Generalized Poisson-based model is determined by the type-II maximum likelihood estimation of model hyperparameters.

We provide in Table 10 details on the number of training epochs (and their corresponding execution times) as executed in an HP Enterprise XL170r E5-2650v4 CPU with 128 GB of RAM memory.

We observe that all models reach convergence within a few number of training epochs (average of 11 epochs for the proposed model). The execution-time overhead incurred by the proposed model results from the aforementioned computation of the normalizing constant in Equation (10). Based on our vectorized implementation, we did not find significant accuracy/execution-time benefits beyond $M = 500$, $s_{max} = 10$, and $x_{max} = 1000$ in the presented real-data experiments.

Improving or boosting the implementation of the proposed model, both via optimized numerical computation of the normalizing constant and its parallelized/distributed training, was out of the scope of this work.

Table 10: Real-world dataset: Training procedure comparison for the studied models.

| Model | Number of epochs | Execution time (s) |
|----------------|------------------------|-----------------------------|
| CNN | 4.600 (\pm 0.800) | 6.642 (\pm 1.660) |
| LSTM | 42.400 (\pm 13.094) | 295.732 (\pm 104.709) |
| RNN | 25.400 (\pm 13.336) | 74.296 (\pm 35.354) |
| Poisson model | 5.000 (\pm 0.632) | 393.072 (\pm 69.404) |
| Proposed model | 11.000 (\pm 1.265) | 36093.951 (\pm 3804.733) |